UNIT IV: Machine Translation and Semantic Processing

Introduction to Machine Translation (MT), Language Divergence and Typology in MT Encoder-Decoder Model for Machine Translation, Translating in Low-Resource Scenarios, MT Evaluation Metrics and Techniques, Bias and Ethical Issues in NLP and Machine Translation, Semantic Analysis and First-Order Logic in NLP, Thematic Roles and Selectional Restrictions in Semantics, Word Senses and Relations Between Senses

MACHINE TRANSLATION AND SEMANTIC PROCESSING

- **Machine Translation** = Automatic cross-lingual text conversion.
- **Semantic Processing** = Understanding & representing meaning in language.
- In NLP, **semantic understanding is critical to high-quality MT**, along with other applications like QA, IR, and chatbots.

Machine Translation in NLP

Machine Translation (MT) is the task of **automatically converting text or speech from one natural language (source) into another (target)** using computational techniques.

Approaches to MT

1. **Rule-Based Machine Translation (RBMT)**
    - Relies on grammar rules, morphology, and bilingual dictionaries.
    - Works well for morphologically complex languages.
    - Example: Early systems like SYSTRAN.
2. **Statistical Machine Translation (SMT)**
    - Uses probabilities derived from bilingual corpora.
    - Example: Phrase-based models.
    - Limitation: Struggles with long-range dependencies and fluency.
3. **Neural Machine Translation (NMT)**
    - Employs deep learning (RNNs, LSTMs, Transformers).
    - Captures context and semantics effectively.

- Current systems: Google Translate, DeepL.

4. **Hybrid MT**
    - Combines rule-based, statistical, and neural methods for better accuracy.

Challenges in MT

- **Ambiguity:** Words with multiple meanings.
- **Idioms & Metaphors:** Hard to translate literally.
- **Word Order Differences:** Languages differ syntactically (e.g., English SVO vs Hindi SOV).
- **Low-Resource Languages:** Lack of large corpora.
- **Domain Adaptation:** A model trained on news text may fail in medical/technical domains.

Semantic Processing in NLP

Semantic Processing refers to techniques for **analyzing, representing, and understanding the meaning** of natural language.

Core Areas of Semantic Processing

1. **Lexical Semantics**
    - Word meaning and relations (synonymy, antonymy, hyponymy, polysemy).
    - Example: *big* ≈ *large*, *rose* → hyponym of *flower*.
2. **Word Sense Disambiguation (WSD)**
    - Selecting the correct sense of a word based on context.
    - Example: *bank* → riverbank vs financial institution.
3. **Semantic Role Labeling (SRL)**
    - Identifies semantic roles in a sentence.
    - Example: *Mary (agent) gave (action) John (recipient) a book (theme).*
4. **Compositional Semantics**
    - Meaning of a sentence derived from the meanings of words and their structure.

5. **Distributional Semantics**
    - Word meanings derived from usage patterns.
    - Implemented via **embeddings** (Word2Vec, GloVe, BERT).
6. **Semantic Networks & Ontologies**
    - Structured representations of concepts and relationships.
    - Examples: WordNet, ConceptNet, DBpedia.

Applications of Semantic Processing

- Improves **Machine Translation** by resolving ambiguities.
- **Question Answering Systems** (understanding intent).
- **Information Retrieval** (context-based search).
- **Chatbots & Dialogue Systems** (semantic understanding).
- **Text Summarization & Categorization.**

MT & Semantic Processing Relationship in NLP

- MT depends heavily on semantic processing.
- Example: Translating *"He went to the bank"* → needs **WSD** to decide if *bank* means financial institution or riverbank.
- Advanced **NMT systems (e.g., Transformer-based models)** leverage **semantic embeddings** (BERT, GPT) for accurate translation.

INTRODUCTION TO MACHINE TRANSLATION (MT)

Machine Translation (MT) is a core task in **Natural Language Processing (NLP)** that focuses on the **automatic translation of text or speech from one natural language (source language) into another (target language)** using computational methods.

Machine Translation is about teaching computers to "understand" and "translate" human languages automatically, with accuracy and fluency.

UNIT IV

The main goal of MT is to make information accessible across languages by removing the **language barrier**. Unlike human translators, MT systems rely on **linguistic rules, statistical models, or neural networks** to generate translations.

Key Features of MT

- Converts source language into target language without human intervention.
- Preserves **meaning** and **context**, not just words.
- Requires understanding of **syntax, semantics, morphology, and pragmatics** of languages.

Evolution of MT

1. **Rule-Based MT (RBMT):** Early systems using grammar rules and dictionaries.
2. **Statistical MT (SMT):** Based on probability models trained on bilingual corpora.
3. **Neural MT (NMT):** Uses deep learning models (e.g., sequence-to-sequence, Transformers) to capture context and semantics.

Applications of MT

- Online translators (Google Translate, DeepL).
- Multilingual communication in business and education.
- Cross-language information retrieval.
- Real-time speech translation (e.g., Skype, Zoom interpreters).

Challenges in MT

- **Ambiguity:** Words with multiple meanings.
- **Idiomatic Expressions:** Hard to translate literally.
- **Word Order Differences:** Languages vary syntactically.
- **Low-Resource Languages:** Lack of sufficient parallel corpora.

LANGUAGE DIVERGENCE AND TYPOLOGY IN MT ENCODER-DECODER MODEL FOR MACHINE TRANSLATION

- **Language Divergence** = structural/semantic differences across languages.

- **Typology** = systematic classification of these differences.
- The **Encoder–Decoder model (with attention/transformers)** in MT is powerful because it can **encode source meaning independent of order/morphology** and **decode appropriately** according to the target language's typological structure.

1. Language Divergence in MT

- **Definition:** Language divergence refers to the **differences across languages** in terms of grammar, morphology, word order, and semantic representation.
- **Examples of Divergence:**
  - **Word Order:** English (SVO: *She eats apples*) vs Hindi (SOV: *Woh seb khati hai*).
  - **Morphology:** English has limited inflections (*walk, walks*), while Turkish has rich morphology (*evlerinizden* = "from your houses").
  - **Lexical Gaps:** Some words/phrases in one language have no direct equivalent in another (e.g., German *Schadenfreude*).

– In MT, these divergences create **challenges for direct word-to-word translation**, requiring models to capture deeper syntactic and semantic structures.

2. Typology in MT

- **Definition:** Linguistic typology classifies languages based on **structural features** like word order, morphology, and syntax.
- **Importance for MT:**
  - Helps design better **pre-processing** and **alignment strategies**.
  - Guides model architecture when dealing with **morphologically rich or free-word-order languages**.
  - Provides insights for **transfer learning** across related languages.

3. Encoder–Decoder Model in MT

The **Encoder–Decoder framework** (foundation for Neural MT) is designed to handle these divergences.

- **Encoder:** Processes the **source language sentence** into a fixed-length or contextual representation (vector).

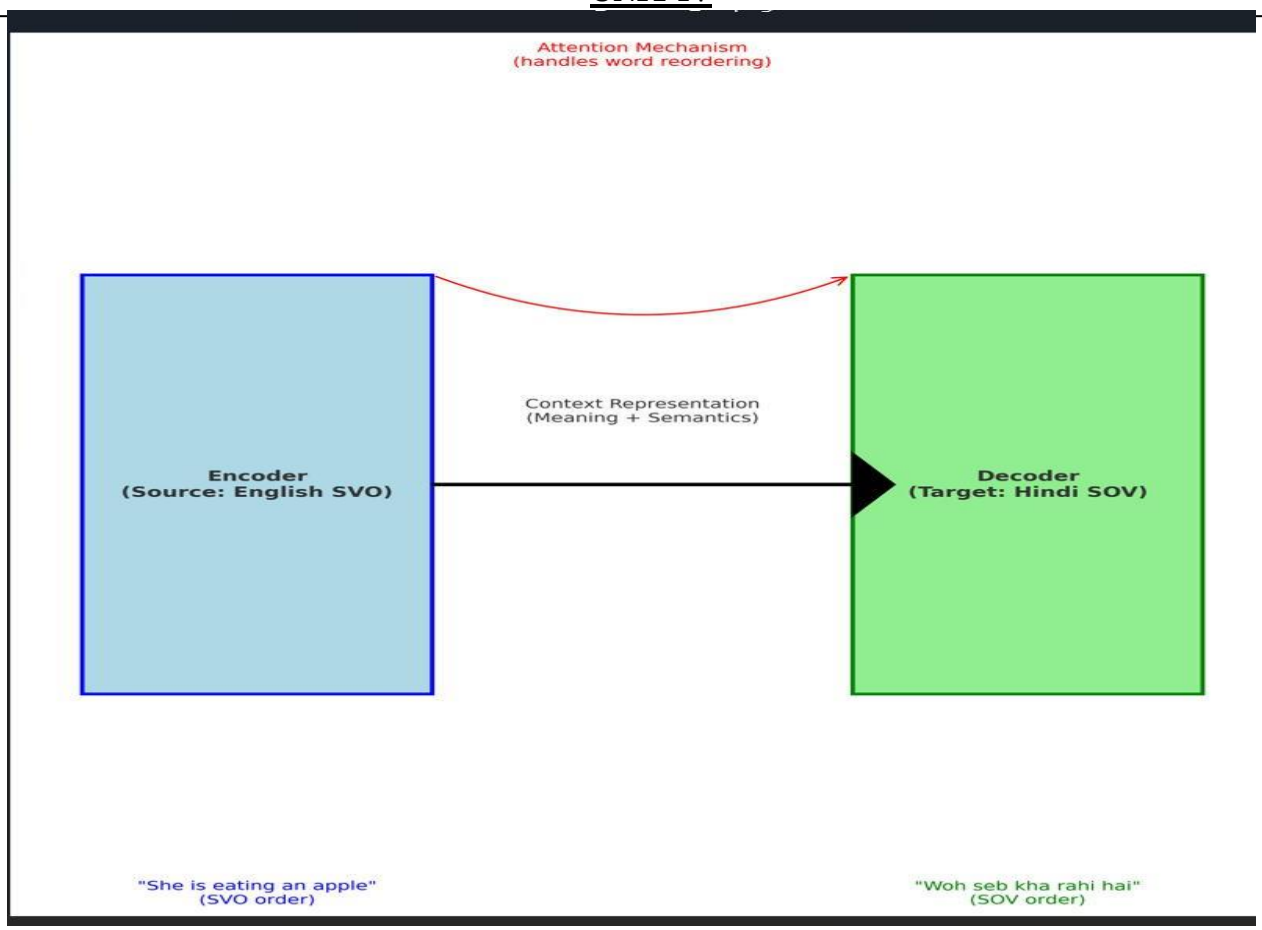
- **Decoder:** Generates the **target language sentence** step by step, conditioned on the encoded representation.
- **Attention Mechanism:** Helps the decoder focus on relevant parts of the source sentence during translation, especially important for long or structurally divergent sentences.

4. Role of Encoder–Decoder in Handling Divergence & Typology

- **Word Order Divergence:**
  - The encoder captures entire sentence context.
  - The decoder reorders words according to the target language syntax (e.g., English → Hindi).
- **Morphological Divergence:**
  - Encoders with **subword/byte-pair encoding (BPE)** handle complex word forms.
  - Useful for agglutinative languages like Finnish, Tamil, Turkish.
- **Semantic Divergence:**
  - Contextual embeddings (e.g., in Transformer-based models like BERT, GPT, mBART) ensure **meaning preservation**.
- **Typological Awareness:**
  - Multilingual NMT models benefit from typology by **sharing representations** across related languages (e.g., Romance languages: Spanish, Italian, French).
  - For highly divergent pairs (English–Chinese), the encoder–decoder needs stronger attention and larger training corpora.

5. Example

- English sentence: *She is eating an apple.* (SVO)
- Hindi translation: *Woh seb kha rahi hai.* (SOV)
- The **encoder** captures meaning in English order.
- The **decoder** rearranges words into Hindi order, respecting its typology.



TRANSLATING IN LOW-RESOURCE SCENARIOS

Translating in low-resource scenarios requires **creativity in data generation (back-translation, augmentation)** and **leveraging multilingual/transfer learning models** to compensate for the lack of parallel corpora.

1. What are Low-Resource Scenarios?

- A **low-resource language** is one that **lacks sufficient parallel corpora, linguistic resources, or digital text data** for training robust Machine Translation (MT) systems.
- Examples: Many African (Yoruba, Amharic), Indian (Kannada, Manipuri), and indigenous languages.
- **Problem:** Most MT approaches (especially Neural MT) require **millions of sentence pairs**, but low-resource languages often have only a few thousand.

2. Challenges in Low-Resource MT

- **Data Scarcity:** Very few parallel corpora for training.
- **Domain Mismatch:** Available data may not represent real-world usage.
- **Morphological Richness:** Low-resource languages are often highly inflected.
- **Divergence:** Structural and typological differences from high-resource languages.
- **Bias & Quality Issues:** Crowdsourced or automatically generated data may contain errors.

3. Approaches to Handle Low-Resource Translation

(a) Data Augmentation

- **Back-Translation:** Translate monolingual target text back into the source language to create synthetic parallel data.
- **Forward Translation:** Translate source monolingual data into target language using weak models.
- **Noise Injection & Paraphrasing:** Generate variations to expand limited data.

(b) Transfer Learning

- **Multilingual NMT (mNMT):** Train one model on multiple related languages so low-resource languages benefit from high-resource ones.
- **Zero-Shot Translation:** Use multilingual models to translate between two languages without direct training pairs (e.g., Hindi ↔ Tamil via English).

(c) Unsupervised MT

- Uses only **monolingual corpora** in both languages with techniques like:
    - Shared word embeddings (cross-lingual representations).
    - Iterative back-translation.

(d) Pretrained Language Models

- Use **large multilingual models** (e.g., mBART, XLM-R, mT5).
- Fine-tune them on small parallel datasets for the target low-resource language.

(e) Pivot-Based Translation

- Translate low-resource → high-resource → target.
- Example: Telugu → English → French.

(f) Community & Crowdsourcing

- Collecting translations from native speakers (e.g., via Wikipedia, local projects).

4. Real-World Examples

- **Google Translate** supports many low-resource languages using multilingual NMT.
- **Masakhane Project (Africa):** Community-driven NMT for African languages.
- **IndicNLP Project:** Building resources for Indian low-resource languages.

5. Summary

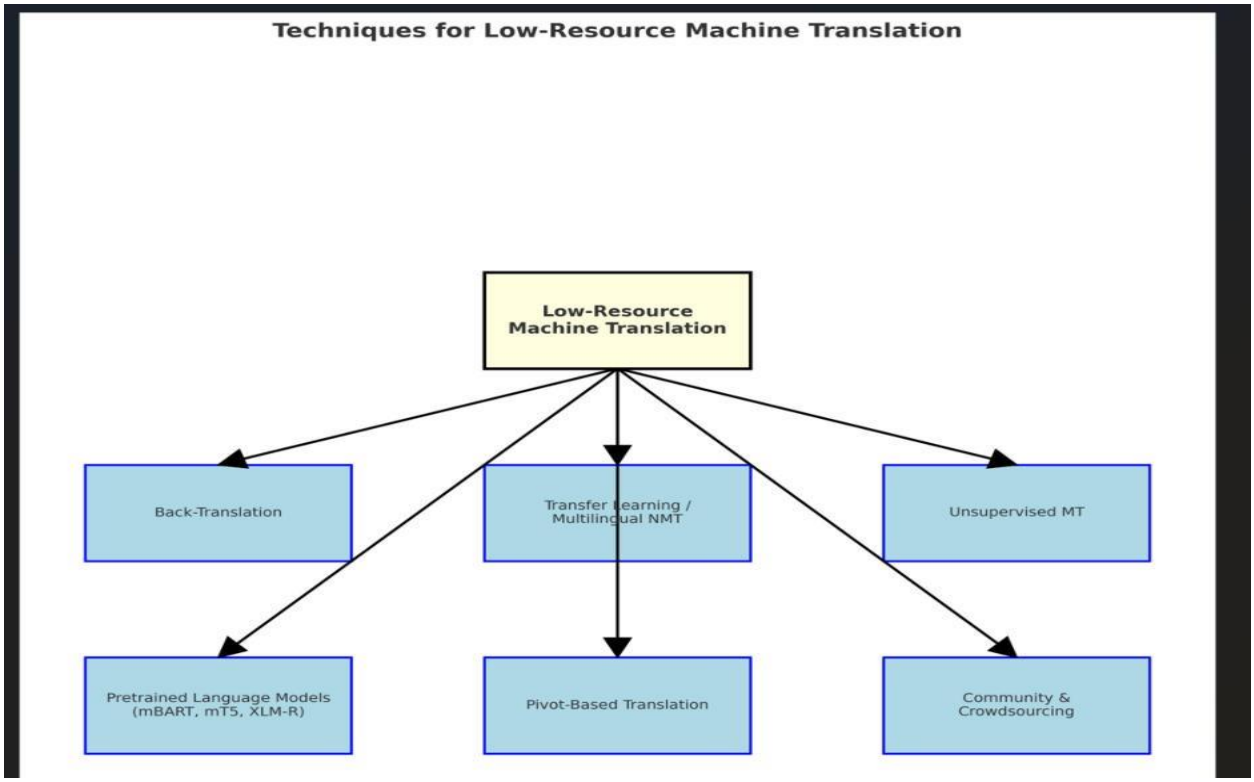
- **Low-resource translation** is a major challenge in NLP.
- Traditional NMT fails due to data scarcity.
- Modern strategies (transfer learning, back-translation, multilingual pretraining, pivoting) help overcome these limitations.
- Success depends not only on algorithms but also on **building more linguistic resources**.

High-Resource vs Low-Resource Machine Translation

- High-resource → straightforward supervised training.
- Low-resource → needs **creative approaches** like augmentation, transfer, and multilingual pretraining.

Aspect	**High-Resource Languages**	**Low-Resource Languages**
Data Availability	Large parallel corpora (millions of sentence pairs)	Very limited parallel corpora (few thousand or none)
Typical Approaches	Supervised Neural MT (Transformer, seq2seq) trained directly on parallel data	Transfer learning, multilingual NMT, pivoting, back-translation, unsupervised MT

Aspect	**High-Resource Languages**	**Low-Resource Languages**
Model Performance	High accuracy, fluent translations	Lower accuracy, may struggle with grammar and meaning preservation
Training Requirements	Requires massive compute and large-scale datasets	Requires data augmentation, synthetic data generation, and fine-tuning
Examples	English ↔ French, English ↔ German, Chinese ↔ English	Yoruba ↔ English, Telugu ↔ French, Amharic ↔ Swahili
Solutions	Direct end-to-end training with attention/transformer models	Back-translation, multilingual pretraining (mBART, mT5), crowdsourcing resources
Challenges	Handling domain adaptation, long sentences	Data scarcity, morphology, typological divergence, noisy/low-quality resources



UNIT IV
MT EVALUATION METRICS AND TECHNIQUES

1. Introduction

Machine Translation (MT) evaluation is the process of **assessing the quality of translated output** produced by MT systems.
It ensures translations are **accurate, fluent, and meaningful**, similar to human translations.

Two broad categories:

1. **Human Evaluation**
2. **Automatic Evaluation**

2. Human Evaluation Techniques

Human judges evaluate translation quality based on linguistic and semantic criteria.

- **Adequacy:** How well the translation preserves the meaning of the source text.
- **Fluency:** Grammatical correctness and naturalness of the target text.
- **Comprehensibility:** Ease with which a reader understands the translation.
- **Ranking/Pairwise Comparison:** Compare outputs of different systems and rank them.

Pros: Captures nuance, idiomatic correctness.
Cons: Time-consuming, costly, subjective.

3. Automatic Evaluation Metrics

Automatic metrics compare MT output against **reference translations** or evaluate quality without references.

(a) Reference-Based Metrics

1. **BLEU (Bilingual Evaluation Understudy):**
    - Compares n-grams (word sequences) between system output and reference translation.
    - Score: 0 (worst) → 1 (best).
    - Widely used, but doesn't capture meaning fully.
2. **NIST:**
    - Extension of BLEU; gives higher weight to *informative* n-grams.

3. **METEOR:**
    - Considers synonyms, stemming, and paraphrases.
    - More correlated with human judgment than BLEU.
4. **TER (Translation Edit Rate):**
    - Measures number of edits (insertions, deletions, substitutions, shifts) needed to transform MT output into reference translation.

(b) Semantic & Embedding-Based Metrics

1. **BERTScore:**
    - Uses contextual embeddings (BERT) to compare semantic similarity.
    - Better at capturing meaning than BLEU.
2. **MoverScore, COMET, BLEURT:**
    - Neural-based evaluation with strong correlation to human judgment.

(c) Reference-Free (Quality Estimation, QE)

- Predicts translation quality **without reference translations**.
- Uses machine learning to model adequacy/fluency directly.
- Useful for low-resource languages where references are scarce.

4. Comparative Summary

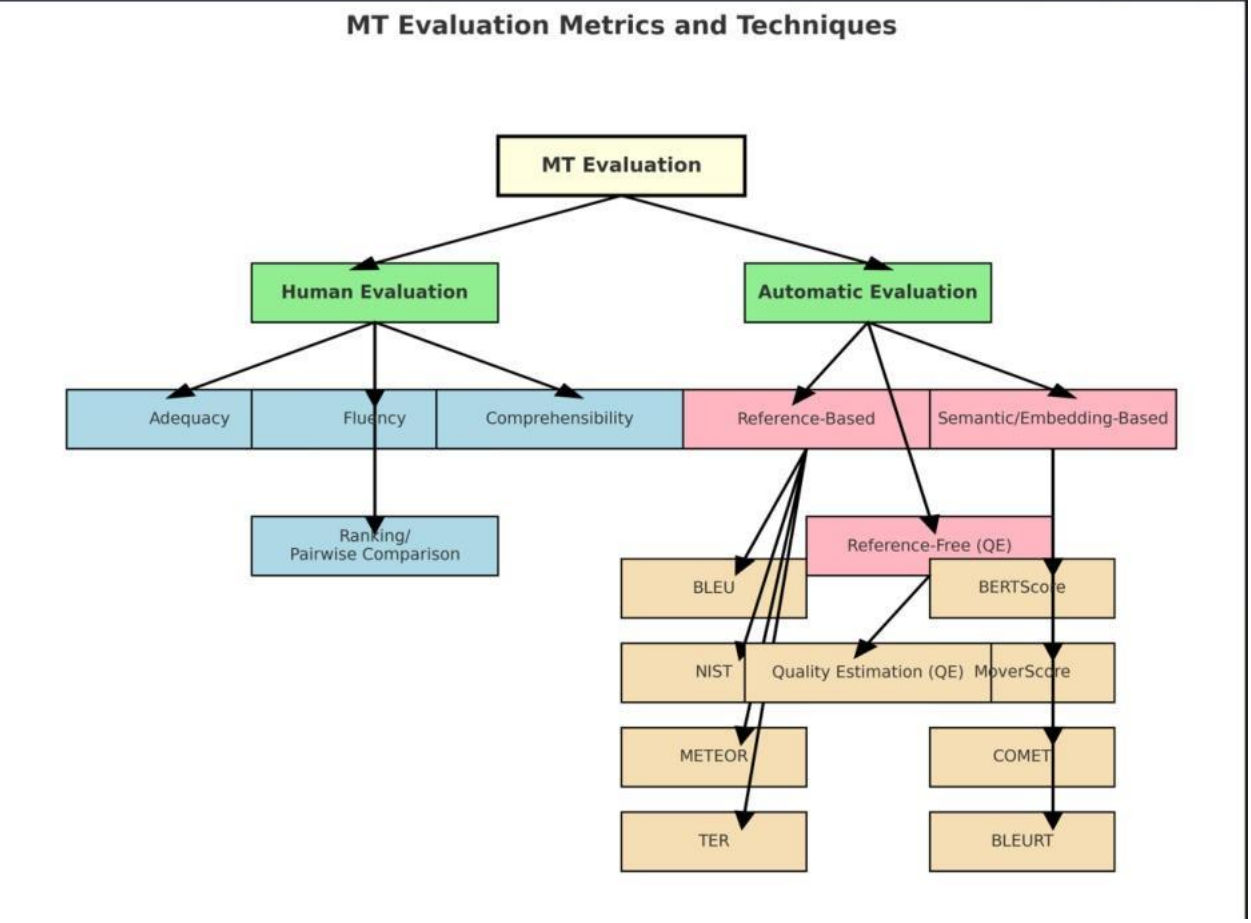
Metric/Technique	**Type**	**Strengths**	**Weaknesses**
Human Evaluation	Manual	Nuanced, captures context	Costly, subjective
BLEU	Automatic	Fast, standard, n-gram based	Ignores meaning, synonyms
NIST	Automatic	Considers informative n-grams	Similar limitations to BLEU
METEOR	Automatic	Synonyms, stemming, paraphrasing	Slower, language-dependent

Metric/Technique	**Type**	**Strengths**	**Weaknesses**
TER	Automatic	Measures effort to correct	May not capture fluency
BERTScore	Automatic	Semantic, contextual	Requires pretrained models
COMET / BLEURT	Automatic	Strong correlation with human judgment	Computationally expensive
QE (Quality Estimation)	Automatic, no reference	Works for low-resource	Less reliable without large training data

5. Conclusion

- **Human evaluation** = gold standard, but costly.
- **Automatic metrics** = fast, objective, scalable.
- Trend: Moving from **surface-level metrics (BLEU, METEOR)** to **semantic/neural metrics (BERTScore, COMET)** for better correlation with human judgment.

MT evaluation combines **human judgment** and **automatic metrics** (BLEU, METEOR, BERTScore, etc.) to ensure translations are both **accurate and fluent**.



BIAS AND ETHICAL ISSUES IN NLP AND MACHINE TRANSLATION

1. Introduction

- NLP and MT systems are widely used (e.g., Google Translate, chatbots, voice assistants).
- However, since they are trained on **large-scale data** collected from the internet, they often **inherit and amplify human biases** present in the data.
- These biases lead to **ethical challenges** like unfair treatment, stereotypes, and misinformation.

2. Sources of Bias in NLP & MT

1. **Data Bias:**
    - Training corpora may overrepresent some languages, cultures, or social groups.

- Example: More data for English and French → better translation quality than for Telugu or Yoruba.

2. **Representation Bias:**
   - Word embeddings capture stereotypes.
   - Example: *doctor → he, nurse → she* associations.
3. **Algorithmic Bias:**
   - Models may optimize for accuracy but ignore fairness.
   - Example: MT translating gender-neutral Turkish sentence *"O bir doktor" → "He is a doctor"* (incorrect gender assumption).
4. **Socio-Cultural Bias:**
   - Systems may fail with culturally specific idioms or dialects.

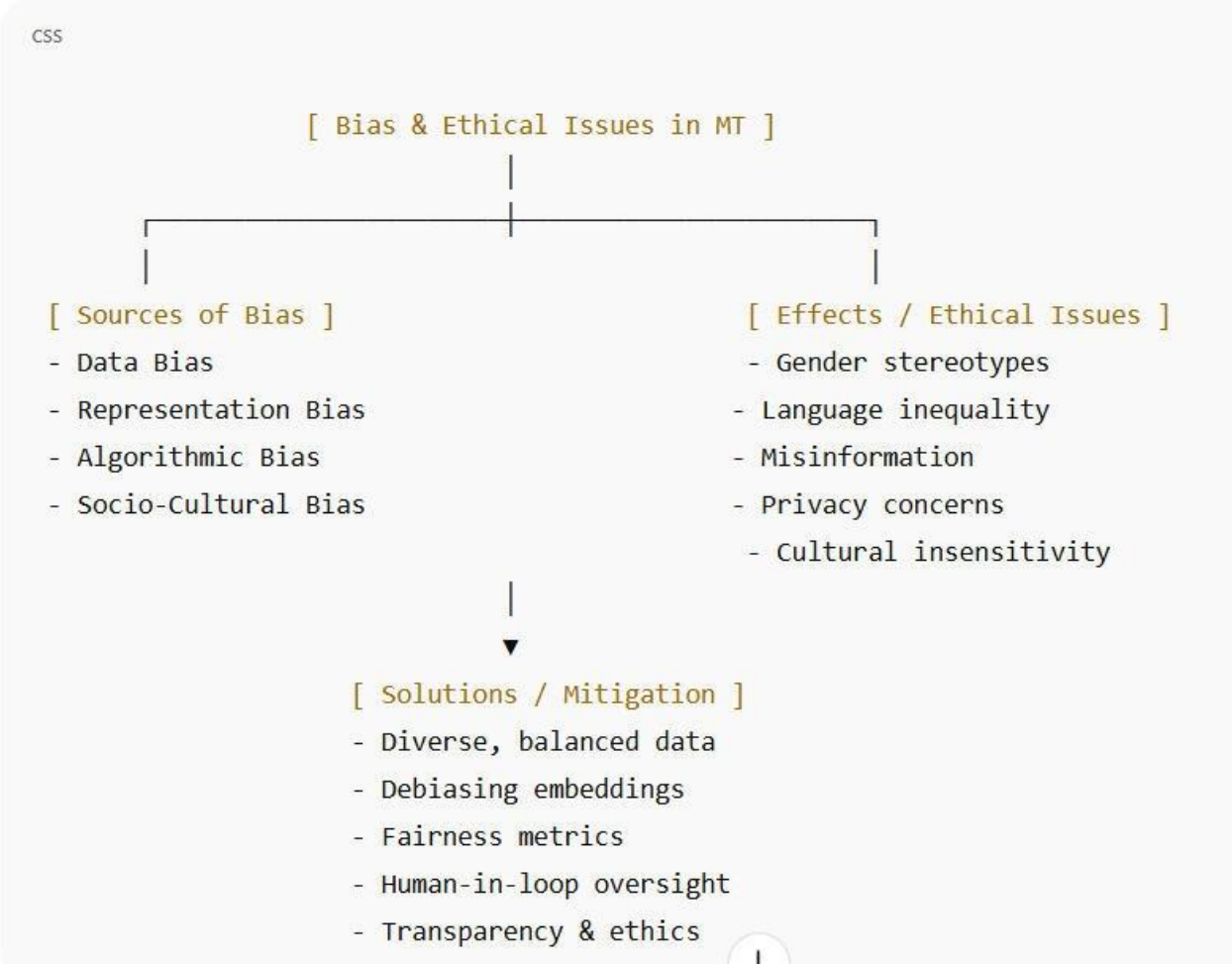
3. Ethical Issues in MT and NLP

1. **Gender & Stereotype Reinforcement:**
   - Translations can reinforce gender roles (*engineer → he*, *teacher → she*).
2. **Language Inequality (Digital Divide):**
   - High-resource languages get accurate MT; low-resource ones are neglected.
   - Leads to exclusion of communities from digital participation.
3. **Misinformation & Trustworthiness:**
   - Poor translations in healthcare/legal contexts may cause harm.
4. **Privacy Concerns:**
   - Using personal text for training without consent can violate privacy.
5. **Cultural Insensitivity:**
   - Incorrect translations of sensitive terms can cause offense or misunderstanding.
6. **Accountability & Transparency:**
   - Users may not know why a system made a biased or wrong translation.

4. Mitigation Strategies

- **Bias-Aware Data Collection:** Balanced, diverse corpora covering different dialects, genders, and contexts.
- **Debiasing Techniques in Embeddings:** Modify word embeddings (e.g., gender-neutral embeddings).
- **Fairness Metrics in MT:** Evaluate translations not only for accuracy (BLEU) but also for fairness and neutrality.
- **Human-in-the-Loop Systems:** Involve human review for sensitive domains (healthcare, law).
- **Transparency & Explainability:** Use Explainable AI (XAI) to show why a translation was produced.
- **Ethical Guidelines & Policies:** Industry and academia must enforce ethical standards.
- **Bias and ethics** in NLP/MT are critical because models influence billions of users daily.
- Problems include **gender stereotypes, language inequality, cultural insensitivity, and privacy risks**.
- Solutions require **better data practices, fairness-aware algorithms, and human oversight** to build **trustworthy and inclusive MT systems**.

Bias in MT arises from **data, algorithms, and cultural factors**, leading to ethical risks like **stereotyping, inequality, and misinformation**.
Ethical MT requires **fair data, debiasing methods, transparency, and accountability**.



SEMANTIC ANALYSIS AND FIRST-ORDER LOGIC IN NLP

- **Semantic Analysis** → Extracts **meaning** from text.
- **FOL** → Provides a **formal, logic-based way** to represent that meaning for reasoning.

Semantic Analysis is about understanding the **meaning** of text beyond its structure. It deals with mapping natural language into **machine-readable representations**.

Key Tasks in Semantic Analysis:

1. **Word Sense Disambiguation (WSD):** Resolving meaning of words in context
    - *Example:* "bank" → river bank vs financial bank
2. **Semantic Role Labeling (SRL):** Identifying roles in sentences
    - *Example:* "John ate an apple" → John = Agent, Apple = Patient

3. **Thematic Relations:** Relations like *cause, effect, instrument, location*
4. **Semantic Parsing:** Converting sentences into formal structures (logic-based or graph-based)
5. **Compositional Semantics:** Meaning of a sentence = combination of meanings of its words + grammar

First-Order Logic (FOL) in NLP

FOL provides a **formal representation of meaning** in natural language using:

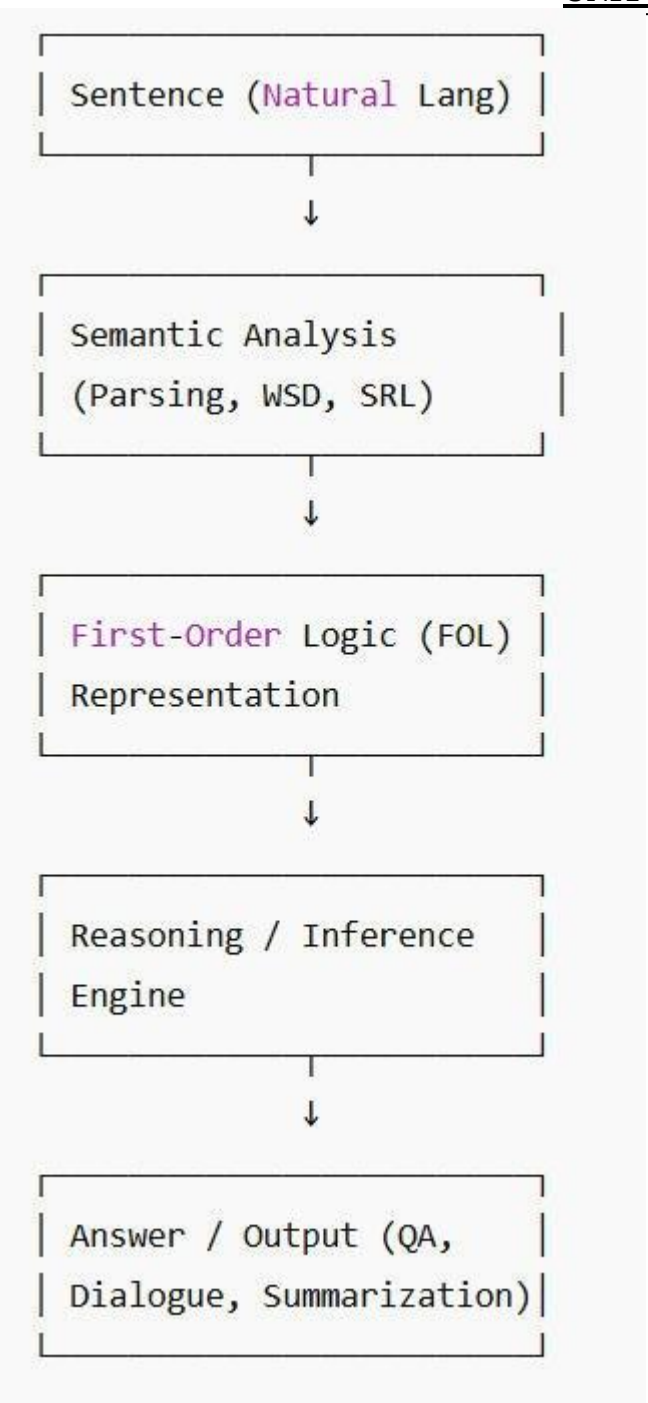
- **Constants:** Specific objects (e.g., *John, Apple*)
- **Predicates:** Properties or relations (e.g., *Loves(John, Mary)*)
- **Quantifiers:**
  - ∀ (for all) → Universal quantification
  - ∃ (there exists) → Existential quantification

Example Conversions:

- Sentence: *"All humans are mortal."*
  FOL: ∀x [Human(x) → Mortal(x)]
- Sentence: *"Some cats are black."*
  FOL: ∃x [Cat(x) A Black(x)]
- Sentence: *"John loves Mary."*
  FOL: Loves(John, Mary)

Applications in NLP

- **Question Answering (QA):** Convert question → FOL → match with KB facts
- **Information Retrieval:** Logical forms help filter relevant facts
- **Text Summarization:** Identify semantic structures and relations
- **Dialogue Systems:** Represent intents and reasoning in logic form



THEMATIC ROLES AND SELECTIONAL RESTRICTIONS IN SEMANTICS

- **Thematic roles** assign **semantic functions** to sentence participants.
- **Selectional restrictions** impose **plausibility checks** on those roles.

1. Thematic Roles (Semantic Roles)

Thematic roles (also called **semantic roles**) describe how **participants** in an event are related to the **verb (predicate)**.
They capture the **who did what to whom, when, where, and how** in a sentence.

Common Thematic Roles:

- **Agent** – the doer of an action
  Example: John (Agent) opened the door.
- **Experiencer** – entity experiencing a state/emotion
  Example: Mary (Experiencer) felt happy.
- **Theme/Patient** – entity affected or moved by an action
  Example: John ate *an apple* (Theme).
- **Instrument** – object used to perform an action
  Example: He cut the bread with *a knife* (Instrument).
- **Beneficiary** – entity for whom an action is performed
  Example: She baked a cake for *her friend* (Beneficiary).
- **Location** – where the action happens
  Example: The party is at *the park* (Location).
- **Goal / Source** – endpoint or starting point of movement
  Example: He traveled *to Paris* (Goal) from *London* (Source).

2. Selectional Restrictions

Selectional restrictions are **semantic constraints** imposed by verbs (predicates) on their arguments.
They ensure that the arguments are **semantically compatible** with the verb.

Examples:

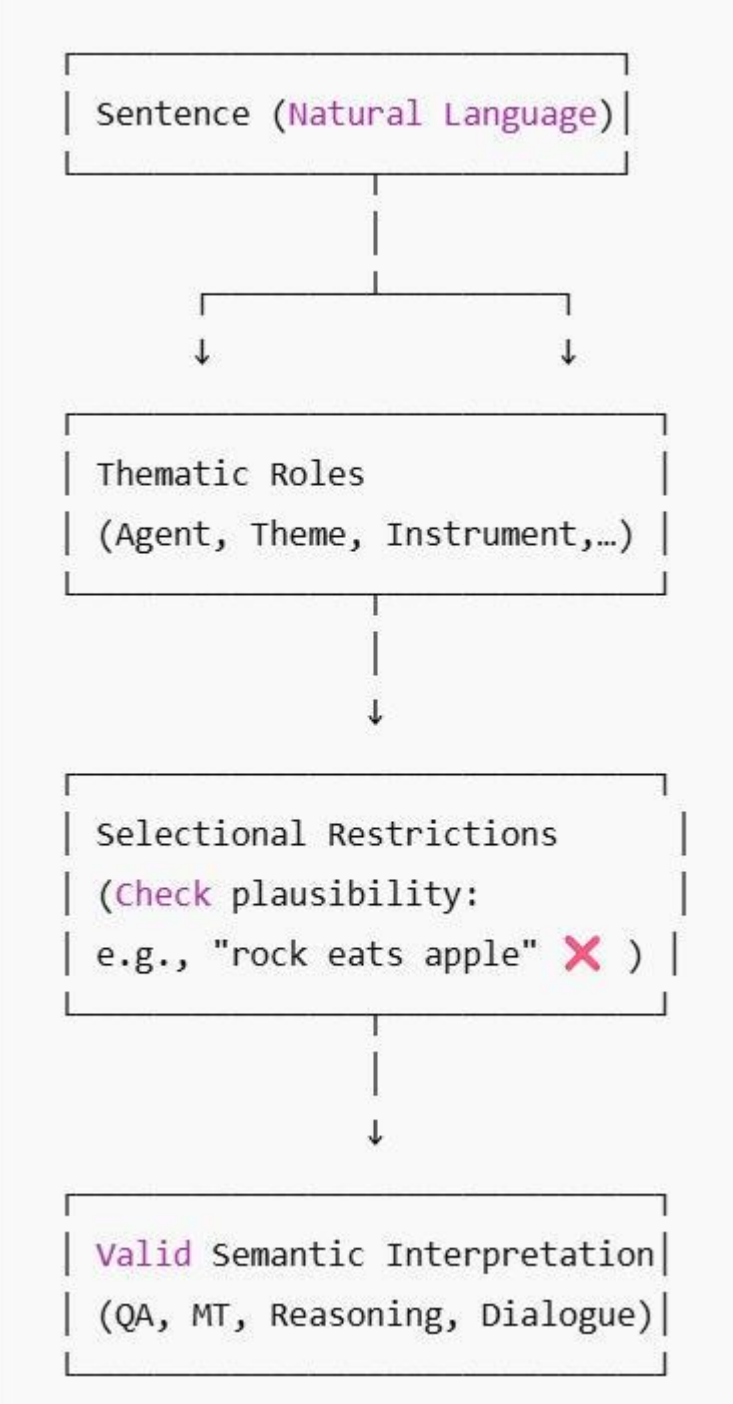
- *"The boy ate an apple."* ✅ (boy = animate Agent, apple = edible Theme)
- *"The rock ate an apple."* ❌ (rock cannot be an animate Agent of "eat")
- *"She drank water."* ✅ (drink requires liquid Theme)
- *"She drank a chair."* ❌ (chair violates selectional restrictions)

Why important?

- Prevents **nonsensical interpretations** in NLP.
- Helps in **Word Sense Disambiguation** (WSD).
- Useful in **semantic parsing, question answering, and MT**.

3. Thematic Roles + Selectional Restrictions in NLP

- Thematic roles help determine **relations between entities**.
- Selectional restrictions **validate plausibility** of those relations.
- Together, they guide **semantic interpretation** of natural language.



WORD SENSES AND RELATIONS BETWEEN SENSES

- **Word senses** = different meanings of a word.

- **Relations between senses** = structured links (synonymy, antonymy, hypernymy, etc.) that organize lexical knowledge.

1. Word Senses

- A **word sense** is a specific meaning of a word in a given context.
- Many words are **polysemous** (multiple related senses) or **homonymous** (unrelated senses).

Example:

- Word: **bank**
  - Sense 1: *financial institution* → "I deposited money in the bank."
  - Sense 2: *side of a river* → "They sat on the river bank."

Word Sense Disambiguation (WSD) is the NLP task of identifying which sense of a word is used in context.

2. Relations Between Word Senses

Lexical semantics defines several **semantic relations** between senses, many of which are captured in **WordNet**.

a) Synonymy (same or similar meaning)

- *big* ↔ *large*
- Used in **thesaurus-based IR**.

b) Antonymy (opposite meaning)

- *hot* ↔ *cold*
- Common in **sentiment analysis**.

c) Hyponymy / Hypernymy (IS-A relation)

- Hyponym: *rose* is a type of *flower*.
- Hypernym: *flower* is a super-category of *rose*.
- Used in **taxonomy construction**.

d) Meronymy / Holonymy (Part-Whole relation)

- Meronym: *wheel* is part of a *car*.

- Holonym: *car* has *wheels*.

e) Polysemy (multiple related senses)

- *head* → of a person, a company, a bed.

f) Homonymy (different, unrelated senses)

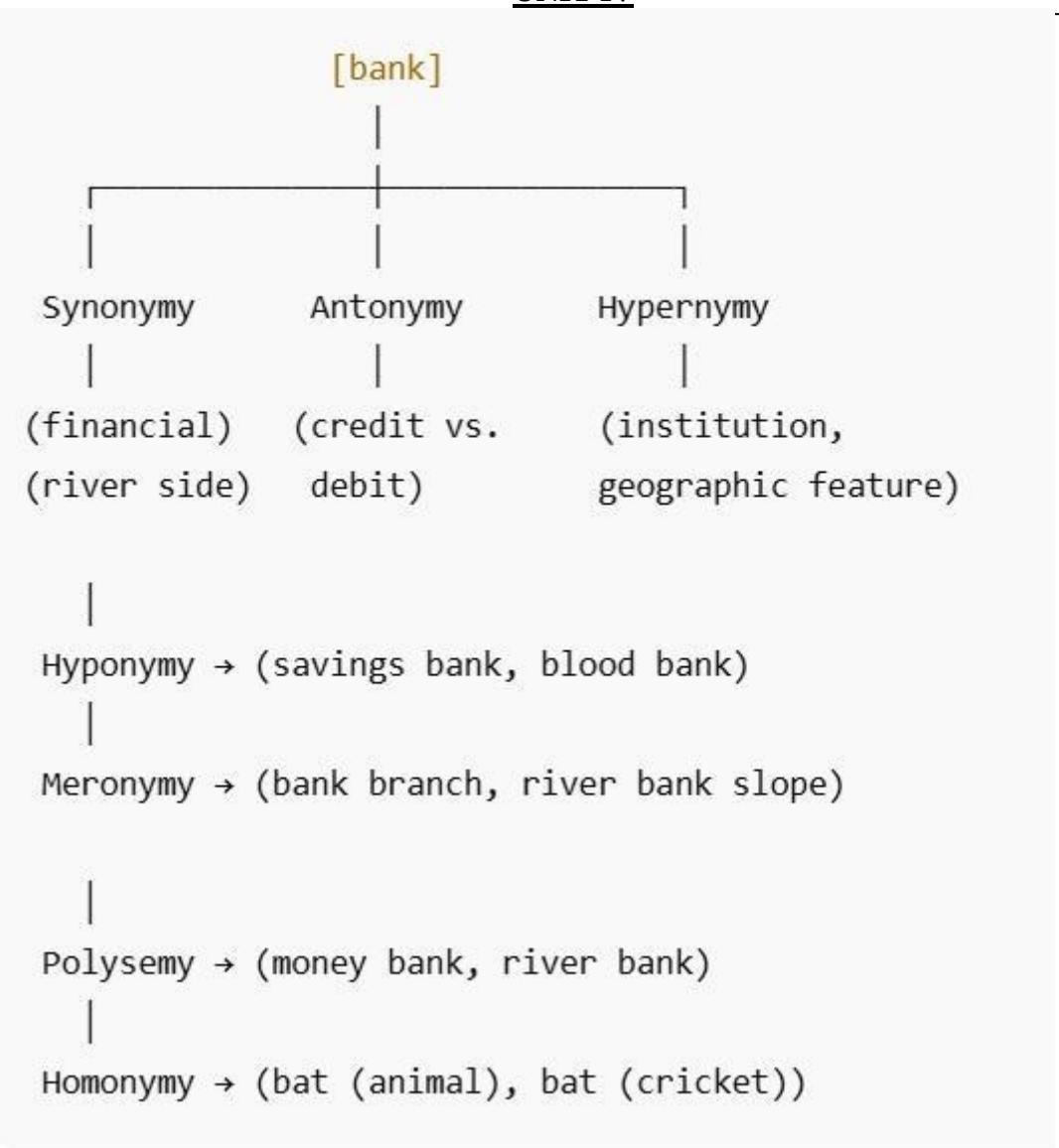
- *bat* → an animal, a cricket bat.

g) Troponymy (manner relation between verbs)

- *walk* ↔ *stroll* (to walk in a leisurely manner).

3. Importance in NLP

- **WSD**: Disambiguating polysemy in translation, QA.
- **MT**: Correct sense selection → accurate translation.
- **IR/Search**: Synonymy and hypernymy expand queries.
- **Ontologies**: Relations form the basis of semantic networks.


[bank]
Synonymy
Antonymy
Hypernymy
(financial)
(river side)
(credit vs.
debit)
(institution,
geographic feature)
Hyponymy → (savings bank, blood bank)
Meronymy → (bank branch, river bank slope)
Polysemy → (money bank, river bank)
Homonymy → (bat (animal), bat (cricket))